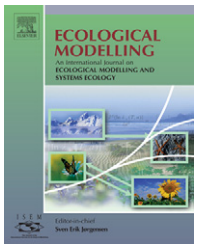
available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



Predicting tree species presence and basal area in Utah: A comparison of stochastic gradient boosting, generalized additive models, and tree-based methods

Gretchen G. Moisen[a,*], *Elizabeth A. Freeman*[a], *Jock A. Blackard*[a], *Tracey S. Frescino*[a], *Niklaus E. Zimmermann*[b], *Thomas C. Edwards Jr.*[c]

[a] *USDA Forest Service, Rocky Mountain Research Station, 507 25th Street, Ogden, UT 84401, USA*
[b] *Department of Landscape Research, Swiss Federal Research Institute WSL Zuercherstrasse 111, CH-8903 Birmensdorf, Switzerland*
[c] *USGS Utah Cooperative Fish and Wildlife Research Unit, College of Natural Resources, Utah State University, Logan, UT 84322-5290, USA*

ARTICLE INFO

Article history:
Accepted 31 May 2006
Published on line 24 July 2006

Keywords:
Species presence
Predictive mapping
Forest inventory
GAMs
Classification trees
Regression trees
See5
Cubist
Stochastic gradient boosting

ABSTRACT

Many efforts are underway to produce broad-scale forest attribute maps by modelling forest class and structure variables collected in forest inventories as functions of satellite-based and biophysical information. Typically, variants of classification and regression trees implemented in Rulequest's© See5 and Cubist (for binary and continuous responses, respectively) are the tools of choice in many of these applications. These tools are widely used in large remote sensing applications, but are not easily interpretable, do not have ties with survey estimation methods, and use proprietary unpublished algorithms. Consequently, three alternative modelling techniques were compared for mapping presence and basal area of 13 species located in the mountain ranges of Utah, USA. The modelling techniques compared included the widely used See5/Cubist, generalized additive models (GAMs), and stochastic gradient boosting (SGB). Model performance was evaluated using independent test data sets. Evaluation criteria for mapping species presence included specificity, sensitivity, Kappa, and area under the curve (AUC). Evaluation criteria for the continuous basal area variables included correlation and relative mean squared error. For predicting species presence (setting thresholds to maximize Kappa), SGB had higher values for the majority of the species for specificity and Kappa, while GAMs had higher values for the majority of the species for sensitivity. In evaluating resultant AUC values, GAM and/or SGB models had significantly better results than the See5 models where significant differences could be detected between models. For nine out of 13 species, basal area prediction results for all modelling techniques were poor (correlations less than 0.5 and relative mean squared errors greater than 0.8), but SGB provided the most stable predictions in these instances. SGB and Cubist performed equally well for modelling basal area for three species with moderate prediction success, while all three modelling tools produced comparably good predictions (correlation of 0.68 and relative mean squared error of 0.56) for one species.

* *Corresponding author*. Tel.: +1 801 625 5384; fax: +1 801 625 5723.
E-mail address: gmoisen@fs.fed.us (G.G. Moisen).

0304-3800/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2006.05.021

1. Introduction

Maps of tree species presence and silvicultural metrics like basal area[1] are needed throughout the world for a wide variety of forest land management applications. These map estimates provide land managers with additional stand-level descriptors needed in support of their decision-making processes. Knowledge of the probable location of certain key species of interest as well as their spatial patterns and associations to other species are vital components of any realistic land management activity. Also needed are spatial representations of relative abundances of these species in the form of basal area predictions which serve as surrogates for important forest attributes like volume and biomass.

In many broad-scale mapping efforts, forest class and structure variables collected in forest inventories are modelled as functions of satellite-based and biophysical information. Examples in the USA include national mapping projects of the U.S. Forest Service, Forest Inventory and Analysis Program[2] (FIA; Bechtold and Patterson, 2005), the U.S.G.S. National Land Cover Data (NLCD; Vogelmann et al., 2001), and the multi-agency LANDFIRE[3] project (Rollins et al., in press). FIA conducts inventories of status and trends in forested ecosystems throughout the U.S. and produces nationwide maps of forest attributes at 250 m resolution. NLCD is a project sponsored by the Multi-Resolution Land Characteristics (MRLC[4]) consortium, to produce nationwide maps of land cover at 30 m resolution. LANDFIRE is an interagency wildland fire, ecosystem, and fuel-mapping project designed to generate 30 m resolution maps of vegetation, fire, and fuel characteristics across the USA.

In all three of these U.S. mapping applications, Rulequest's©[5] See5 and Cubist software packages are the tools being used for modelling and prediction. These variants on classification and regression trees are partially described in Quinlan (1986, 1993). A few of the reasons behind See5 and Cubist's popularity in broad-scale modelling exercises are: their relative ease of use; their fast model-building behavior for either continuous or discrete response variables; their lack of distributional assumptions; and their ability to generate relatively good model predictions in a production environment. While these automated modelling tools are widely used in large remote sensing applications, they are not easily interpretable and also use model fitting algorithms that are proprietary and thus not published or known to the user.

Numerous other techniques have been used for predictive mapping in forestry applications including nearest neighbor methods (e.g., Tomppo, 1991; Franco-Lopez et al., 2001; Ohmann and Gregory, 2002; McRoberts et al., 2002), multivariate adaptive regression splines (e.g., Iverson and Prasad, 2001; Moisen and Frescino, 2002; Prasad and Iverson, 2002), random forests (e.g. Bunn et al., 2005; Prasad et al., 2006), and artificial neural networks (e.g., Foody et al., 2003; Thuiller, 2003), to name just a few. Generalized additive models (GAMs) (Hastie and Tibshirani, 1986, 1990) are widely used in the ecological literature (Guisan et al., 2002), and offer advantages of interpretability and reliable predictions. In addition, stochastic gradient boosting (SGB; Friedman, 2001, 2002) is a recent advance in predictive modeling, but has yet to be tested for predicting species distributions.

In this paper, we compared the predictive performances of three modelling techniques (See5/Cubist, GAMs and SGBs) for mapping species presence and basal area of 13 tree species in the mountains of Utah, U.S.A. Models were constructed using forest inventory field data and ancillary topographic and satellite-based information. For species presence/absence, prediction accuracies were evaluated using traditional threshold-dependent measures of accuracy, threshold independent Receiver Operating Characteristic (ROC) plots and associated Areas Under the Curve (AUC). For species basal area, the effects of modelling techniques on global measures of accuracy and residual plots were assessed. For both species presences and basal areas, accuracy measures and analyses were conducted using independent test sets.

2. Materials and methods

2.1. Data description

2.1.1. Study region

The study area comprises over 6 million hectares (ha) of land predominantly in Utah, with small portions overlapping into Wyoming and Idaho in the Interior West region of the USA (Fig. 1). The area is made up of two ecological provinces (Bailey et al., 1994) that include the Wasatch and Uinta Mountain Ranges in the north, and a series of high plateaus in the south. The study area is delineated by the United States Geological Survey (USGS) zone 16 (Fig. 1), as defined by the national mapping protocols within the MRLC project (Homer and Gallant, 2001). Our analyses are restricted to forested lands, comprising approximately half the area of zone 16. This zone consists of heterogeneous mountainous terrain reaching elevations of over 3000 m, and includes a wide variety of vegetation types ranging from sagebrush shrub-steppe through conifer-dominated forests to alpine communities.

2.1.2. Response variables

FIA provided data for the response variables of species presence and basal area. A network of permanent sample plots has been established across the country at an intensity of approximately one plot per 2400 ha, and data collection is conducted under an annual rotating panel system (Bechtold and Patterson, 2005). Sample plots in the study region have been measured since 1993. A systematic sample of field plots was originally established on a 2.5 km grid on lands administered by the National Forests, and on a 5 km grid across all other land ownerships. Under the annual rotating panel, 1/10 of the plots established on the 5 km grid are revisited each year. Of the 3456 plots available in zone 16, only forested and single-condition plots were used in these analyses. Thus, we restricted our

[1] Sum of cross-sectional areas of tree stems measured at 1.4 m above ground, expressed per land unit area.
[2] http://www.fia.fs.fed.us/.
[3] http://www.landfire.gov.
[4] http://www.mrlc.gov.
[5] www.rulequest.com.

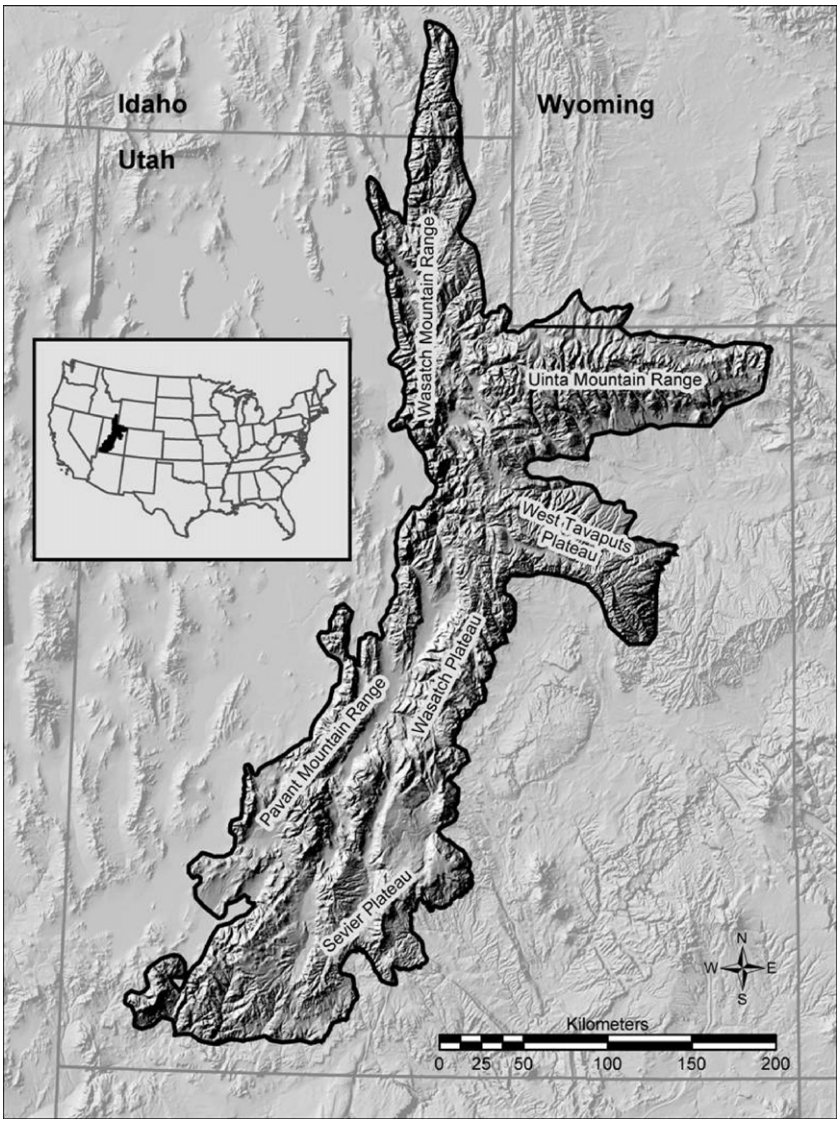


Fig. 1 – The study area is delineated by the United States Geological Survey zone 16 (Homer and Gallant, 2001), as defined by the Multi-Resolution Land Characteristics consortium.

analyses to forested plots falling completely within a particular forest condition (dictated by tree size, composition, and density), and removed plots that had complete or partial non-forest vegetation. A total of 1930 sample plots remained for our analyses, 20% of which were withheld as an independent test set. At each sample plot, data on forest stand structure were sampled in four circular sub-plots, one center and three satellite sub-plots regularly spaced around the center sub-plot (Fig. 2). Each sub-plot covers a radius of 7.3 m and the total area represented by one sample plot was approximately 0.6 ha. At each FIA forested sample plot, extensive stand- and tree-level measurements were collected. Individual tree measurements were compiled and combined with stand-level variables to produce plot-level summaries. A particular tree species was considered present if at least one tree 2.54 cm or greater in diameter (measured at breast height, ~1.4 m) of that species was tallied. Otherwise the species was considered absent. In addition, total tree basal area by each of 13 tree species was compiled for each plot. Table 1 summarizes the number of occupied plots and prevalence for each of 13 tree species in zone 16.

2.1.3. *Predictor variables*

Satellite imagery from the MRLC consortium comprised some of the predictor variables used in these analyses. The MRLC consortium was organized to acquire and process Landsat 7 Enhanced Thematic Mapper Plus (ETM+) data for multiple dates across the United States and Puerto Rico and to coordinate efforts to produce the NLCD for 2001 using this imagery and ancillary data.[6] Satellite imagery was collected for three different time periods representing the temporal dynamics of vegetation: early (spring), peak (summer), and late growing seasons (fall). The steps used for standardizing the imagery

[6] http://www.mrlc.gov.

Table 1 – Number of occupied plots and prevalence for each of the 13 most common tree species in Zone 16

Latin name	Symbol	Common name	Prevalence	Plots w/species present
Abies concolor	ABCO	White fir	0.12	233
Abies lasiocarpa	ABLA	Subalpine fir	0.22	429
Acer grandidentatum	ACGR3	Bigtooth maple	0.06	119
Cercocarpus ledifolius	CELE3	Curlleaf mountain-mahogany	0.08	147
Juniperus osteosperma	JUOS	Utah juniper	0.25	473
Juniperus scopulorum	JUSC2	Rocky Mountain juniper	0.12	230
Pinus contorta	PICO	Lodgepole pine	0.12	230
Pinus edulis	PIED	Common or twoneedle pinyon	0.21	405
Picea engelmannii	PIEN	Englemann spruce	0.18	357
Pinus ponderosa	PIPO	Ponderosa pine	0.09	173
Populus tremuloides	POTR5	Quaking aspen	0.32	623
Pseudotsuga menziesii	PSME	Douglas-fir	0.22	417
Quercus gambelii	QUGA	Gambel oak	0.14	273
Total number of forested plots				1930
Total number of plots				2997

processing are summarized in the Landsat science data users handbook (Irish, 2001), in Homer et al. (2004), and in the literature cited therein. Here, we briefly discuss the basic processing steps applied to the data set we used in our study area.

Landsat scenes are processed in several steps in order to produce consistent and reliable spectral information of land cover. These steps include geometric and terrain correction, and a suite of radiometric corrections. Geometric and terrain corrections rectify the image to match features on the ground, removing distortion caused by the sensor and the topography on the ground. Radiometric correction calibrates the raw digital numbers (DN, 0–255) recorded by the sensor with the reflectance on the ground. This removes distortions caused by the sensor, sun illumination geography, and atmospheric effects, the latter being the most challenging. Several atmospheric correction algorithms have been developed (e.g. Liang et al., 1997), but for large area applications, many users are still concerned with possible unknown errors that may arise due to uncertainties in the ground and atmospheric data necessary to run these algorithms (e.g. Cohen et al., 2001). The alternative is to convert DN values to at-satellite reflectance values, thereby normalizing the illumination geometry, yet reducing the need for atmospheric correction (Huang et al., 2002).

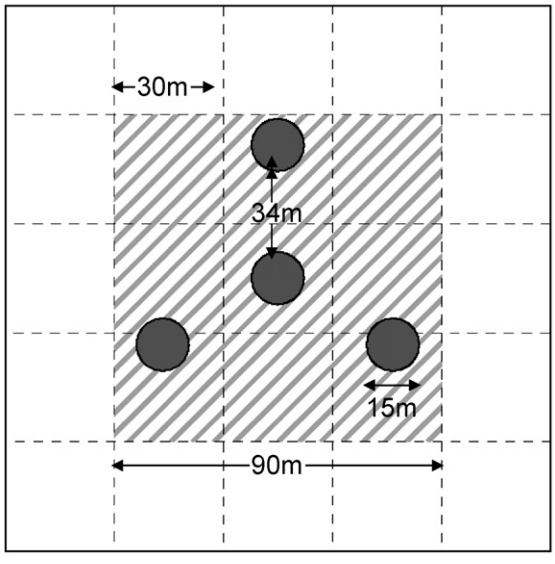


Fig. 2 – Four subplots comprising an FIA sample plot. The gridlines indicate the approximate Landsat Thematic Mapper (TM) pixels of roughly 30 m side length. The highlighted 3 × 3 pixels represent the 90 m × 90 m area re-sampled for linking the FIA plot data with the TM-based predictors.

The MRLC ETM+ bands were first resampled using cubic convolution into an Albers Equal Area map projection. Then all bands were georectified using cubic convolution re-sampling techniques, terrain corrected using the USGS national elevation dataset, and the DN values were converted to at-sensor radiance (L) values. Registration accuracy standards were within 1 pixel (30 m) root mean square error (RMSE). Next, bands 1–5 and 7 were converted to at-sensor reflectance (ρ) values. Then, the tasseled cap (TC) transformation (Kauth and Thomas, 1976; Christ and Cicone, 1984) – a linear recombination of bands 1–5 and 7 – was applied according to Huang et al. (2002) resulting in three new products, namely the soil brightness index (SBI), the green vegetation index (GVI) and the wetness index (WI). The TC transformation is often applied to regions where a full atmospheric correction is not feasible. It typically explains up to 95% of the variance per scene. Additionally, the normalized difference vegetation index (NDVI) was calculated using bands 3 (RED) and 4 (NIR), so that NDVI = (NIR + RED)/(NIR-RED). All of these indices were calculated based on ρ-values per band. Finally, L of band 6 (thermal infrared) was converted to at-satellite temperature (B9 hereafter), providing a physically based variable.

The indices, along with bands 1–5 and 7, were resampled to a 90 m cell size based on a 3 × 3 moving window using the "focalmean" and "focalstd" GRID functions in ESRI ArcGIS©. This was done to ensure coverage of an area that is at least the full spatial extent of the dependent forest inventory plot data (Fig. 2), which is considerably larger than one 30 m Landsat TM pixel.

In addition to the ETM+ bands and indices, we used location (in x and y coordinates in Albers Equal Area map projection), elevation (m) from the National Elevation Dataset (NED; Gesch

Table 2 – Description of predictor variables

Type	Names	Description
Location	AlbX	Albers equal area X coordinate
	AlbY	Albers equal area Y coordinate
Topography	Elev	Elevation in meters
	Slope	Slope in degrees
	Asp-E	"Eastness" [sine(aspect)]
	Asp-N	"Northness" [cosine(aspect)]
Reflectance	B1spr.M, B1sum.M, B1fal.M B1spr.S, B1sum.S, B1 fal.S	Band 1: Visible blue [0.450–0.515 μm]
	B2spr.M, B2sum.M, B2 fal.M B2spr.S, B2 sum.S, B2 fal.S	Band 2: Visible green [0.525–0.605 μm]
	B3spr.M, B3sum.M, B3 fal.M B3spr.S, B3 sum.S, B3 fal.S	Band 3: Visible red [0.630–0.690 μm]
	B4spr.M, B4sum.M, B4 fal.M B4spr.S, B4 sum.S, B4 fal.S	Band 4: Near infrared [0.775–0.900 μm]
	B5spr.M, B5sum.M, B5 fal.M B5spr.S, B5 sum.S, B5 fal.S	Band 5: Shortwave infrared [1.550–1.750 μm]
	B7spr.M, B7sum.M, B7 fal.M B7spr.S, B7 sum.S, B7 fal.S	Band 7: Shortwave infrared [2.090–2.350 μm]
	B9spr.M, B9sum.M, B9 fal.M B9spr.S, B9 sum.S, B9 fal.S	"Band 9": At sensor surface temperature
Indices	TC1spr.M, TC1sum.M, TC1fal.M TC1spr.S, TC1sum.S, TC1fal.S	Tasseled cap soil brightness index [using B1–B5, B7]
	TC2spr.M, TC2sum.M, TC2fal.M TC2spr.S, TC2sum.S, TC2fal.S	Tasseled cap green vegetation index [using B1–B5, B7]
	TC3spr.M, TC3sum.M, TC3fal.M TC3spr.S, TC3sum.S, TC3fal.S	Tasseled cap wetness index [using B1–B5, B7]
	NDVIspr.M, NDVIsum.M, NDVIfal.M NDVIspr.S, NDVIsum.S, NDVIfal.S	Normalized difference vegetation index [using B3, B4]
	B9spr.M, B9sum.M, B9fal.M B9spr.S, B9sum.S, B9fal.S	"Band 9": at sensor surface temperature [using B6H]

Variable names with "spr", "sum", and "fal" refer to spring, summer and fall imagery respectively, while ".M" and ".S" refer to mean and standard deviations obtained when indices and reflectance bands were resampled to 90 m based on 3 × 3 moving windows.

et al., 2002) and slope and aspect values derived from the NED by their respective GRID functions in ArcGIS©. These variables were also resampled to 90 m. Aspect took the form of two transformed variables "northness" and "eastness" (Clark et al., 1999) where northness is the cosine of aspect and eastness is the sine of aspect. Digital values of these predictor layers were extracted from imagery and DEM data for each FIA location. Table 2 summarizes all predictor variables.

2.2. *Modelling*

Species presence was modelled for each species individually. FIA plots in the training data set where the species of interest was present were given a response value of one, and all other plots were given a value of zero. Predictions in the form of probability of species presence were made for all plots in the test set. Species basal area was modeled using only those FIA plots in the training data set where the species of interest was present. Predictions of basal area for a particular species were made for those plots in the independent test set having that species present.

2.2.1. *Generalized additive models*

GAMs (Hastie and Tibshirani, 1986, 1990) are nonparametric extensions of generalized linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989) which are in turn an extension of the classical linear model. GAMs have been used extensively in ecological applications and model fitting details are well documented (see Yee and Mitchell, 1991). An extensive review of GAM applications in ecology is given in Guisan et al. (2002). Predictive mapping using GLMs and GAMs pertaining to forest inventory applications is illustrated in Moisen and Edwards (1999), Frescino et al. (2001), and Moisen and Frescino (2002). Muñoz and Felicisimo (2004) compare GLMs to alternative predictive mapping methods. Spatial prediction of species distributions using GAMs is explored by Austin (2002), while a GAM application tool for ecological analyses and spatial prediction has been built by Lehmann et al. (2002). Recent advances in GAM methodology have been made by Wood and Augustin (2002), as well as Yee and MacKenzie (2002).

GAMs were fit in R (R Development Core Team, 2005; Ihaka and Gentleman, 1996) using the *gam* package (Hastie and

Tibshirani, 1996; Venables and Ripley, 2002). The binary species presence/absences were modeled using a binomial family with logit link, while species basal area was modeled with a gamma family and log link. For both continuous and discrete responses, predictor variables first entered the models individually using a smoothing spline with a relatively conservative smoothing parameter to avoid fitting noise. Because of the unwieldy number of predictor variables, a series of stepwise procedures were applied using each of seven classes of predictor variables individually to reduce the number of predictor variables and to determine if those that remain should enter with smooth or linear terms. These seven classes of predictor variables included location and topography, spring bands, spring indices, summer bands, summer indices, fall bands, and fall indices. Stepwise results were compiled to create unique models for each species. An alternative approach to creating subsets of predictor variables is offered by Leathwick et al. (2006).

2.2.2. *Variants on classification and regression trees*

Classification and regression trees, also known as recursive partitioning regression (Breiman et al., 1984), are widely used in remote sensing applications (e.g. Friedl and Brodley, 1997; Friedl et al., 1999; Hansen et al., 2000; Huang and Townshend, 2003; Prasad and Iverson, 2002; Schwarz and Zimmermann, 2005). Trees subdivide the space spanned by the predictor variables into regions for which the values of the response variable are approximately equal, and then estimate the response variable by a constant in each of these regions. The tree is called a classification tree if the response variable is qualitative, and a regression tree if the response variable is quantitative. Two recent enhancements to tree-based methods have met with considerable success in mapping applications (Chan et al., 2001). One is known as bagging, or bootstrap aggregation (Bauer and Kohavi, 1998; Breiman, 1996). The other is called boosting (Freund and Schapire, 1996) with its variant Resampling and Combining (ARCing) (Breiman, 1998). These iterative schemes each produce a committee of expert tree models by resampling with replacement from the initial data set. Afterwards, the expert tree models are averaged using a plurality voting scheme if the response is discrete, or simple averaging if the response is continuous. The difference between bagging and ARCing/boosting is the way in which data are resampled. In the former, all observations have equal probability of entering the next bootstrap sample; while in the latter, problematic observations (i.e. observations that have been frequently misclassified) have a higher probability of selection.

The species presence models were generated using classification trees with boosting implemented in Rulequest's© See5 software package. Boosting with ten trials and pruning were the two options used. Basal area models were constructed through Rulequest's© Cubist software package, a proprietary variant on regression trees with piecewise non-overlapping regression. Specific software options used in our study region included ten committee models, use of rules alone, minimum rule cover of 1% of cases, extrapolation up to 10%, with no maximum number of rules.

2.2.3. *Stochastic gradient boosting*

Stochastic gradient boosting (Friedman, 2001, 2002) is related to both boosting and bagging. Many small classification or regression trees are built sequentially from "pseudo"-residuals (the gradient of the loss function of the previous tree). At each iteration, a tree is built from a random sub-sample of the dataset (selected without replacement) producing an incremental improvement in the model. Using only a fraction of the training data increases both the computation speed and the prediction accuracy, while also helping to avoid over-fitting the data. An advantage of stochastic gradient boosting is that it is not necessary to pre-select or transform predictor variables. It is also resistant to outliers, as the steepest gradient algorithm emphasizes points that are close to their correct classification.

To date, there have been very few published ecological applications of stochastic gradient boosting. Lawrence et al. (2004) illustrate the use of stochastic gradient boosting as a refinement of classification tree analysis in a remote sensing problem. Application in other fields includes discrimination of freshwater residency in a coastal fishery from scales collected from subadult fish (Cappo et al., 2005), microscopy image analysis of bread (Lindgren and Rousu, 2002), graphical estimation of a slate deposit (Matias et al., 2004), and calibrating spectroscope measurements of organic chemicals in plant samples (Sheperd et al., 2003).

Stochastic gradient boosting was implemented through the *gbm*[7] package (Ridgeway, 1999) within R. Model fitting options include distribution, interaction depth, bagging fraction, shrinkage rate, and training fraction. Friedman (2001, 2002) and Ridgeway (1999) provide guidelines on appropriate settings for model fitting options. These recommendations coupled with exploratory analyses conducted on an independent data set led to the following option settings. A Bernoulli distribution was used for species presence/absence models, and a Gaussian distribution was used for species basal area models. Interaction depth, which controls the number of nodes in the tree and thus the maximum possible interactions, was set at ten nodes. Bagging fraction controls the fraction of the training data randomly selected for calculating each tree, and was set at 0.3 for these analyses. Shrinkage rate controls the learning speed of the algorithm and we used *gbm*'s default learning rate of 0.001. Training fraction sets aside a portion of the data for computing an out-of-sample estimate of the loss function. As an independent test set had already been established before beginning the modelling exercises, the training fraction was left at its default value of 1.0, and the out-of-bag method was used for determining the optimal number of boosting iterations.

2.3. *Evaluation criteria*

2.3.1. *Presence/absence*

Because the utility of maps for different management applications cannot be captured in a single map accuracy number,

[7] http://www.i-pensieri.com/gregr/gbm.shtml provides a good overview of the development of boosting leading up to stochastic *gbm*.

Table 3 – Accuracy measures computed using independent test data for the 13 tree species and the three presence/absence modelling techniques: Rulequest's© See5 model, general additive models (GAM), and stochastic gradient boosting (SGB)

Species	Kappa			Sensitivity			Specificity		
	See5	GAM	SGB	See5	GAM	SGB	See5	GAM	SGB
ABCO	0.48	0.48	0.57	0.52	0.59	0.57	0.95	0.93	0.96
ABLA	0.53	0.68	0.60	0.65	0.82	0.64	0.89	0.91	0.94
ACGR3	0.38	0.48	0.51	0.41	0.64	0.59	0.97	0.95	0.96
CELE3	0.38	0.46	0.51	0.34	0.38	0.41	0.97	0.99	0.99
JUOS	0.77	0.79	0.81	0.85	0.89	0.96	0.93	0.92	0.90
JUSC2	0.25	0.28	0.33	0.47	0.36	0.58	0.85	0.92	0.85
PICO	0.73	0.77	0.78	0.80	0.86	0.82	0.96	0.96	0.97
PIED	0.63	0.69	0.69	0.77	0.76	0.77	0.88	0.93	0.93
PIEN	0.74	0.76	0.76	0.70	0.83	0.74	0.98	0.96	0.98
PIPO	0.54	0.55	0.52	0.50	0.61	0.53	0.97	0.95	0.96
POTR5	0.66	0.69	0.74	0.74	0.77	0.81	0.92	0.92	0.93
PSME	0.46	0.51	0.51	0.71	0.80	0.69	0.81	0.80	0.86
QUGA	0.50	0.72	0.66	0.53	0.84	0.67	0.94	0.94	0.96

The thresholds were selected to maximize Kappa.

several global measures were used to assess the predictive performance of the models. All measures were constructed using an independent test set created by randomly withholding 20% of the plots from FIA's probability-based sample in the study area. The first three measures are threshold dependent and include sensitivity, specificity, and Kappa (Cohen, 1960). Fielding and Bell (1997) provide a review of these accuracy measures. Sensitivity, or proportion of true positives, reflects a model's ability to detect a presence given a species actually occurs at a location. Specificity, or proportion of true negatives, reflects a model's ability to predict an absence where a species does not exist. Thresholds optimized on these measures can, however, be deceptive when prevalence is very low or very high. The Kappa statistic measures the proportion of correctly classified units after accounting for the probability of chance agreement. Kappa has an advantage over sensitivity and specificity in that it is more resistant to prevalence (Manel et al., 2001), though it still requires a choice of threshold. Thresholds were chosen to optimize Kappa for each species and model, and global measures of model accuracy were computed using the independent test set.

ROC plots provide a threshold independent method of evaluating the performance of presence/absence models. In a ROC plot the true positive rate (sensitivity) is plotted against the false positive rate (1.0, specificity) as the threshold varies from 0 to 1. A good model will achieve a high true positive rate while the false positive rate is still relatively small; thus the ROC plot will rise steeply at the origin, then level off at a value near the maximum of 1. The ROC plot for a poor model (whose predictive ability is the equivalent of random assignment) will lie near the diagonal, where the true positive rate equals the false positive rate for all thresholds. Thus the area under the curve (AUC) is a good measure of overall model performance, with good models having an AUC near 1, while poor models have an AUC near 0.5.

Tests of significant differences between models based on respective AUC were conducted following DeLong et al. (1988) using library functions developed at the Mayo Clinic.[8] These functions provided both an overall test for equality of areas, as well as pair-wise comparisons between each model.

2.3.2. *Continuous response*

For the basal area by species, site-specific measures of accuracy calculated on the same independent test set described above included relative mean squared error (MSE) and correlation. The relative MSE is calculated by dividing the MSE obtained when applying the model in question to the independent test set by the MSE obtained when simply using the sample mean as a prediction for the entire independent test set. Relative MSE's less than 1 indicate improvement using the model in question over a simple sample mean. Values greater than one indicate the models in question perform worse than a simple sample mean. The correlation is the standard measure of the linear relationship between two quantitative variables. Correlation values range from zero to one, with larger values indicating stronger linear relationship between the predicted and observed response.

3. Results

3.1. *Presence/absence*

SGB had higher values for the majority of the species for specificity and Kappa, while GAMs had higher values for the majority of the species for sensitivity (Table 3). Thresholds were chosen to maximize kappa, with different thresholds chosen for each of the three modelling techniques. Fig. 3 illustrates how a threshold that is optimal for one model may be inappropriate for other models. Here, the effect of chang-

[8] Developed by Beth Atkinson and Doug Mahoney, available at http://mayoresearch.mayo.edu/mayo/research/biostat/splusfunctions.cfm for Unix, and http://www.stats.ox.ac.uk/pub/MASS3/Winlibs/ for Windows.

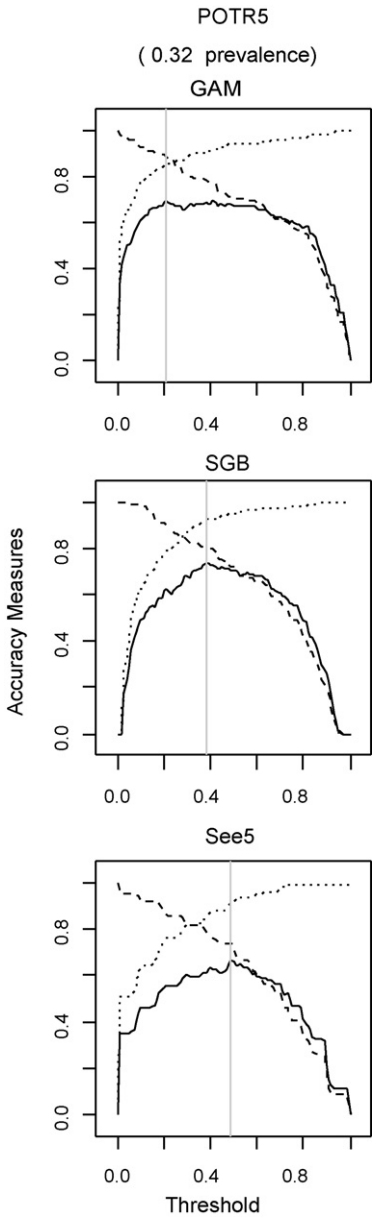


Fig. 3 – Error threshold plots for Presence/absence of *Populus tremuloides* for each of the three models. Different accuracy measures are indicated by different line types: sensitivity, specificity, kappa. The vertical lines represent the threshold that maximizes kappa for the three models.

ing a threshold on sensitivity (dashed line), specificity (dotted line), and kappa (solid line) is shown for *Populus tremuloides* using each of the three modelling techniques. In all cases, as one increases the threshold, model sensitivity decreases while specificity increases. For this species, a threshold chosen to maximize kappa for the GAM model would result in lower than necessary kappa values for the See5 and the SGB models.

ROC plots were studied for each species and modelling technique, with three of the species illustrated in Fig. 4. These species have similar prevalence, illustrating that model performance can vary independently from species prevalence.

In order to test for significant differences between modelling techniques, we began with an overall test of equality of AUC for all the models. Eight species yielded significant *p*-values for the overall tests (Table 4). Identifying where these between-model differences occurred was accomplished by applying a Bonferroni correction to the model-to-model comparisons. Using three pair-wise comparisons, the alpha value was calculated as 0.0167. Resulting *p*-values are also shown in Table 4. Specific model differences are illustrated in Table 5. Here, a line under two model techniques indicates the pair-wise test was unable to distinguish a significant difference between those two models. A line under all three models indicates no significant difference between them. Eight of the 13 species showed an overall significant difference. Of those eight, GAM was significantly highest for one species, and either GAM or SGB were significantly highest for five species. For one species the best model was either SGB or See5, and for the remaining species the pair-wise tests were unable to determine highest AUC value (due to the lower power from the Bonferroni correction).

3.2. *Species basal area results*

Relative MSE and correlations were obtained using independent test data to apply basal area models for all 13 tree species and three modelling techniques (Table 6). Nine out of 13 species had very poor predictive models with relative MSE of 0.8 or higher and correlations 0.5 or below. For those nine marginal species, SGB had the lowest relative MSE's for six of them and highest correlations for seven. Interestingly, the relative MSE from the SGB models never exceeded one, (never predicted worse than the sample mean), for any species. GAM and Cubist, on the other hand, produced relative MSE greater than one on seven and four species respectively.

Better results were obtained for *Pinus contorta*, *Pinus edulis*, and *Quercus gambelii*. For all three of these species, Cubist and SGB produced similar accuracy values, with Cubist slightly better. *Populus tremuloides* was the species for which the highest correlations and lowest relative MSE's were obtained. This was the most prevalent of all the species, and the only deciduous type assuming normal tree form. Here, all three modelling techniques were competitive.

4. Discussion

Maps of tree species presence and basal area are needed for forest management activities, and results from these analyses have implications for species mapping efforts. First, the widely used See5 and Cubist algorithms may not be providing the best predictive power. In comparing performance of modelling techniques, SGB had higher values for the majority of the species for naïve accuracy, specificity and kappa; while GAMs had higher values for a majority of the species for sensitivity. In testing for significant difference between resultant AUC values, eight of the 13 species showed overall significant difference. Of those eight, the GAM was significantly highest for one species, while either GAM or SGB were significantly higher for five species. See5 tied with SGB for only one species, and for the remaining species, the pair-wise tests were unable

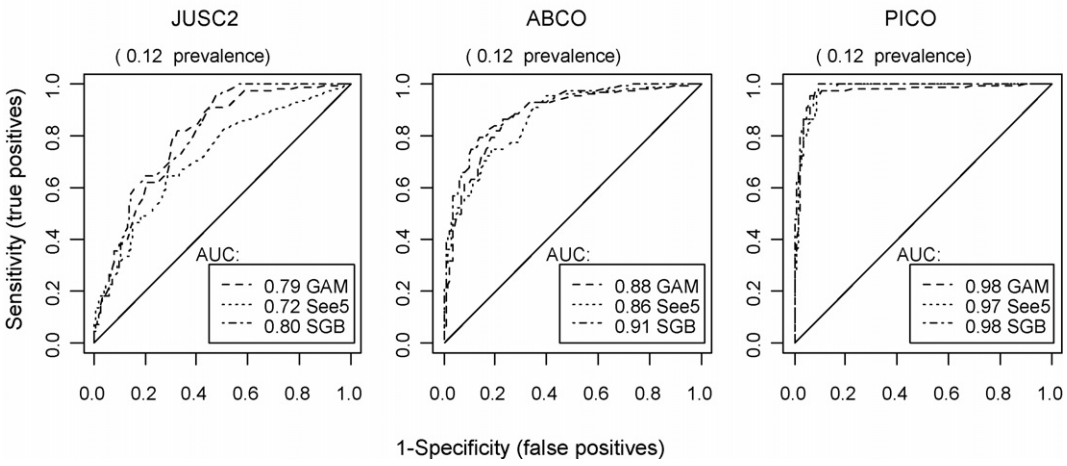


Fig. 4 – Receiver Operating Characteristic (ROC) plots and Area Under the Curve (AUC) for general additive models (GAM), Rulequest's© Cubist model, and stochastic gradient boosting (SGB) models for presence/absence of *Juniperous scopulorum* (JUSC2), *Abies concolor* (ABCO), and *Pinus contorta* (PICO).

to distinguish between models. These results may warrant consideration of alternative modelling techniques within production mapping environments.

A second implication from these analyses is that modelling continuous variables, like individual tree species basal area, is a difficult task. Most results for the GAM and Cubist models were no better than, and sometimes worse, than simply predicting the sample mean (as indicated by a relative MSE less than one.) One clear pattern that emerged was that SGB typically outperformed both Cubist and GAMs, and even for those species for which all models performed badly, SGB never performed worse than using the sample mean. This stability is valuable in any mapping environment.

Of course, there are potential sources of error in the data sets used for modelling, including positional uncertainties, issues pertaining to image processing algorithms, and limitations in field data collection. These errors may be affecting the overall accuracies of the models. Some assumptions related to data locations are that the field sampled response data are accurately georeferenced, that the predictor data are accurately registered to each other and to the ground reference data, and that the 90 m × 90 m focal windows encompass the respective reference locations. The MRLC Landsat products were registered with quality restrictions of less that 1 pixel RMSE. Although these standards are considered high quality for this extensive data set, a shift of one pixel may affect this registration with the ground reference locations.

Regarding the image processing algorithms, standardization methods for the MRLC Landsat product do not eliminate all noise in the data, and may, in fact, introduce errors into the modelling process. For example, a cubic convolution resampling procedure was performed to reproject and

Table 4 – Area under the curve and significance tests for the 13 tree species and the three presence/absence modelling techniques: general additive models (GAM), Rulequest's See5 model, and stochastic gradient boosting (SGB)

Symbol	Area Under the Curve			*p*-Values			
	GAM	See5	SGB	Overall	GAM-See5	GAM-SGB	See5-SGB
ABCO	0.87	0.86	0.91	0.0952			
ABLA	0.92	0.85	0.90	0.0044	0.0012	0.2095	0.0026
ACGR3	0.87	0.87	0.93	0.0014	0.9147	0.0121	0.0712
CELE3	0.88	0.82	0.88	0.3249			
JUOS	0.95	0.96	0.97	0.1456			
JUSC2	0.79	0.72	0.80	0.0136	0.0428	0.7250	0.0035
PICO	0.97	0.97	0.98	0.0165	0.9752	0.1208	0.0282
PIED	0.93	0.91	0.93	0.0238	0.3177	0.5925	0.0075
PIEN	0.96	0.94	0.96	0.1359			
PIPO	0.86	0.89	0.87	0.6148			
POTR5	0.92	0.90	0.94	0.0023	0.2521	0.0831	0.0006
PSME	0.84	0.81	0.85	0.0143	0.1267	0.6126	0.0037
QUGA	0.94	0.85	0.90	0.0001	0.0001	0.0080	0.0366

The overall *p*-value is for the null hypothesis of no difference between the performances of the three models. For species where significant differences were found in the overall test, pair-wise comparisons are given. The pair-wise tests were evaluated using the Bonferroni correction with $\alpha = 0.05/3 = 0.0167$.

Table 5 – Summary of the results of the pair-wise tests of model performance for the three presence/absence modelling techniques: general additive models (GAM), Rulequest's See5 model, and stochastic gradient boosting (SGB)

Species	Area Under the Curve <-- low		high-->
ABLA	See5	SGB	GAM
ACGR3	GAM	See5	SGB
JUSC2	See5	GAM	SGB
PICO	GAM	See5	SGB
PIED	See5	GAM	SGB
POTR5	See5	GAM	SGB
PSME	See5	GAM	SGB
QUGA	See5	SGB	GAM

The underscores connect models where the pair-wise tests were unable to find a significant difference.

georectify the Landsat imagery. This procedure was used to maintain the spatial integrity of the imagery, thereby minimizing inconsistencies among different layers (Homer et al., 2004). Although retaining spatial consistency among layers, this resampling technique alters the spectral integrity of the imagery by interpolating data from the nearest 16 pixel values.

Table 6 – Relative mean square error (MSE) and correlations obtained on independent test data when applying basal area models by the 13 tree species and the three modelling techniques: general additive models (GAM), Rulequest's© Cubist model, and stochastic gradient boosting (SGB)

Species	Relative MSE			Correlation		
	GAM	Cubist	SGB	GAM	Cubist	SGB
ABCO	1.37	0.98	0.93	0.13	0.24	0.31
ABLA	2.04	0.91	0.86	0.19	0.34	0.38
ACGR3	3.33	2.78	1.00	−0.27	−0.06	0.03
CELE3	0.83	1.01	0.97	0.57	0.18	0.19
JUOS	0.87	1.16	0.96	0.40	0.08	0.23
JUSC2	8.18	1.18	1.00	0.04	−0.06	−0.01
PICO	0.98	0.74	0.76	0.42	0.49	0.49
PIED	1.07	0.74	0.81	0.33	0.54	0.58
PIEN	1.14	0.95	0.87	0.30	0.30	0.35
PIPO	1.27	0.90	1.00	0.20	0.33	0.28
POTR5	0.59	0.56	0.60	0.68	0.68	0.68
PSME	1.95	0.83	0.80	0.16	0.43	0.53
QUGA	1.24	0.72	0.80	0.36	0.53	0.50

We also acknowledge that difficulty in predicting basal area may, at least in part, be due to field data collection techniques. FIA plot design may result in basal area data with high variance. The plots may simply be too small for accurate calculation of basal area for areas the size of aggregated pixels.

A final cautionary note is that statistical differences between modelling techniques may not necessarily translate to relevant difference from a management perspective. Conversely, models that did not produce significantly different global performance measures may produce wildly different maps resulting in drastically different implications for management decisions. Ultimately, maps should be evaluated in light of their intended use.

5. Conclusions

Three modelling techniques were compared for predicting species presence and basal area for 13 tree species in the mountains of Utah, USA. Modelling techniques included the widely used variants of classification and regression trees implemented in See5 and Cubist, as well as GAMs, and SGB.

For predicting species presence, SGB had higher values for a majority of the species for specificity and kappa, while GAMs had higher values for a majority of the species for sensitivity. In evaluating resultant AUC values, where significant differences could be detected between models, GAM and/or SGB models had significantly higher results than the See5 models.

Predictions of basal area were poor for nine out of 13 species, having relative MSE's greater than 0.8 and correlations lower than 0.5. SGB, however, provided the most stable predictions for these species in that relative MSE's never exceeded one. For the three species with moderate prediction success, SGB and Cubist were competitive, while all three modelling tools produced comparably good predictions for the one remaining species.

Based on these analyses, the authors suggest the following when mapping species distributions and tree basal area: (1) implement SGB, and possibly GAMs, in a production environment for mapping species distributions as supplements or alternatives to the widely used See5 package; (2) question whether or not modelling individual tree basal area, or other continuous variables, will produce predictions that are better than simply applying a sample mean; (3) explore model performance in terms of implications for management decisions.

Acknowledgments

We acknowledge the staff of the U.S. Forest Service, Forest Inventory and Analysis Program in the Interior West for their contributions in field data collection, compilation, GIS services, administrative support, and financial backing. Thanks also go to the many participants in the 2004 Workshop on Generalized Regression and Spatial Prediction whose ideas and discussions motivated this paper. We are especially grateful to the workshop organizers, A. Guisan, A. Lehmann, J. Overton, S. Ferrier, and R. Aspinall, for providing a wonderful forum for discussion and collaborative research.

REFERENCES

Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol. Model. 157 (2–3), 101–118.

Bailey, R.G., Avers, P.E., King, T., McNab, W.H. (Eds.), 1994. Ecoregions and Subregions of the United States (map). U.S. Geological Survey, Washington, DC. Scale 1:7,500,000, colored, accompanied by a supplementary table of map unit descriptions, prepared for the U.S. Department of Agriculture, Forest Service.

Bauer, E., Kohavi, R., 1998. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach. Learn. 5, 1–38.

Bechtold, W.A., Patterson, P.L. (Eds.), 2005. The enhanced forest inventory and analysis program—national sampling design and estimation procedures. Gen. Tech. Rep. SRS-80. U.S. Department of Agriculture, Forest Service, Southern Research Station, Asheville, NC.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 26, 123–140.

Breiman, L., 1998. Arcing classifiers. Ann. Stat. 26, 801–849.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth and Brooks/Cole, Monterey, CA.

Bunn, A.G., Goetz, S.J., Fiske, G.J., 2005. Observed and predicted responses of plant growth to climate across Canada. Geophys. Res. Lett. 32, L16710.

Cappo, M., De'ath, G., Boyle, S., Aumend, J., Olbrich, R., Hoedt, F., Perna, C., Brunskill, G., 2005. Development of a robust classifier of freshwater residence in barramundi (*Lates calcarifer*) life histories using elemental ratios in scales and boosted regression trees. Marine Freshwater Res. 56 (5), 713–723.

Chan, J.C.-W., Huang, C., DeFries, R.S., 2001. Enhanced algorithm performance for land cover classification using bagging and boosting. IEEE T. Geosci. Remote. 39 (3), 693–695.

Christ, E.P., Cicone, R.C., 1984. A physically-based transformation of Thematic Mapper data—the TM Tasseled Cap. IEEE T. Geosci. Remote. GE-22, 256–263.

Clark, D.B., Palmer, M.W., Clark, D.A., 1999. Edaphic factors and the landscape-scale distributions of tropical rain forest trees. Ecology 80 (8), 2662–2675.

Cohen, J., 1960. A coefficient of agreement of nominal scales. Educ. Psychol. Meas. 20, 37–46.

Cohen, W.B., Maiesperger, T.K., Spies, T.A., Fiorella, M., 2001. Modelling forest cover attributes as continuous variables in a regional context with Thematic Mapper data. Int. J. Remote Sens. 22, 2279–2310.

DeLong, E.R., Delong, D.M., Clarke-Pearson, D.L., 1988. Comparing areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44 (3), 837–845.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24 (1), 38–49.

Foody, G.M., Boyd, D.S., Cutler, M.E.J., 2003. Predictive relations of tropical forest biomass from Landsat TM data and their transferability between regions. Remote Sens. Environ. 85 (4), 463–474.

Franco-Lopez, H., Ek, A.R., Bauer, M.E., 2001. Estimation and mapping of forest stand density, volume, and cover type using the *k*-nearest neighbors method. Remote Sens. Environ. 77, 251–1709.

Frescino, T.S., Edwards Jr., T.C., Moisen, G.G., 2001. Modelling spatially explicit forest structural attributes using generalized additive models. J. Veg. Sci. 12, 15–26.

Freund, Y., Schapire, R., 1996. Experiments with a new boosting algorithm. In: Machine Learning: Proceedings of the 13th International Conference, Morgan Kauffman, San Francisco, pp. 148–156.

Friedl, M.A., Brodley, C.E., 1997. Decision tree classification of land cover from remotely sensed data. Remote Sens. Environ. 61, 399–409.

Friedl, M.A., Brodley, C.E., Strahler, A.H., 1999. Maximizing land cover classification accuracies produced by decision trees at continental to global scales. IEEE T. Geosci. Remote. 37, 969–977.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data An. 38 (4), 367–378.

Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M., Tyler, D., 2002. The national elevation dataset. Photogram. Eng. Rem. S. 68, 5–12.

Guisan, A., Edwards Jr., T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Model. 157, 89–100.

Hansen, M., DeFries, R.S., Townshend, J.R.G., Sohlberg, R., 2000. Global land cover classification at 1 km spatial resolution using a classification tree approach. Int. J. Remote Sens. 21, 1331–1364.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman and Hall, New York.

Hastie, T.J., Tibshirani, R.J., 1986. Generalized additive models. Stat. Sci. 1, 297–318.

Hastie, T.J., Tibshirani, R.J., 1996. S Archive: mda, StatLib (http://lib.stat.cmu.edu/S/).

Homer, C., Gallant, A., 2001. Partitioning the conterminous United States into mapping zones for Landsat TM land cover mapping. USGS Draft White Paper. http://landcover.usgs.gov.

Homer, C., Huang, C.Q., Yang, L.M., Wylie, B., Coan, M., 2004. Development of a 2001 national land-cover database for the United States. Photogram. Eng. Remote Sens. 70 (7), 829–840.

Huang, C., Wylie, B., Yang, L., Homer, C., Zylstra, G., 2002. Derivation of a tasselled cap transformation based on Landsat 7 at-satellite reflectance. Int. J. Remote Sens. 23, 1741–1748.

Huang, C., Townshend, J.R.G., 2003. A stepwise regression tree for nonlinear approximation: applications to estimating subpixel land cover. Int. J. Remote Sens. 24 (1), 75–90.

Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. J. Comput. Graph. Stat. 5, 299–314.

Irish, R.R., 2001. Landsat 7 Science Data User's Handbook, Report 430-15-01-003-0, National Aeronautics and Space Administration. (http://ltpwww.gsfc.nasa.gov/IAS/handbook/handbook_toc.html).

Iverson, L.R., Prasad, A.M., 2001. Potential changes in tree species richness and forest community types following climate change. Ecosystems 4 (3), 186–199.

Kauth, R.J., Thomas, G.S., 1976. The tasseled cap—a graphic description of the spectral-temporal development of agricultural crops as seen in Landsat. In: Proceedings of the Symposium on Machine Processing of Remotely Sensed Data, West Lafayette, IN, June 29–July 1. LARS, Purdue University, West Lafayette, IN, pp. 41–51.

Lawrence, R., Bunn, A., Powell, S., Zambon, M., 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. Remote Sens. Environ. 90, 331–336.

Leathwick, J.R., Elith, J., Hastie, T., 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions 199 (2), 188–196.

Lehmann, A., Overton, J.M., Leathwick, J.R., 2002. GRASP: generalized regression analysis and spatial prediction. Ecol. Model. 157, 189–207.
Liang, S., Fallahdal, H., Kalluri, S., 1997. An operational atmospheric correction algorithm for Landsat Thematic Mapper imagery over the land. J. Geophys. Res. 102, 173–186.
Lindgren, J.T., Rousu, J., 2002. Microscopy image analysis of bread using machine learning methods. Technical Report C-2002-68, Department of Computer Science, University of Helsinki.
Manel, S., Ceri Williams, H., Ormerod, S.J., 2001. Evaluating presence/absence models in ecology: the need to account for prevalence. J. Appl. Ecol. 38, 921–931.
Matias, J.M., Vaamonde, A., Taboada, J., Gonzalez-Manteiga, W., 2004. Support vector machines and gradient boosting for graphical estimation of a slate deposit. Stoch. Environ. Res. Risk A 18 (5), 309–323.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman and Hall, New York.
McRoberts, R.E., Nelson, M.D., Wendt, D.G., 2002. Stratified estimation of forest area using satellite imagery, inventory data, and the *k*-nearest neighbors technique. Remote Sens. Environ. 82, 457–468.
Moisen, G.G., Edwards Jr., T.C., 1999. Use of generalized linear models and digital data in a forest inventory of northern Utah. J. Agric. Biol. Environ. S. 4, 372–390.
Moisen, G.G., Frescino, T.S., 2002. Comparing five modeling ptimisin techniques for predicting forest characteristics. Ecol. Model. 157, 209–225.
Muñoz, J., Felicisimo, A.M., 2004. Comparison of statistical methods commonly used in predictive modelling. J. Veg. Sci. 15, 285–292.
Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized Linear Models. J. R. Stat. Soc. A 135, 370–384.
Ohmann, J.L., Gregory, M.J., 2002. Predictive mapping of forest composition and structure with direct gradient analysis and nearest neighbor imputation in Coastal Oregon, USA. Can. J. Forest Res. 32, 725.
Prasad, A.M., Iverson, L.R., 2002. Predictive vegetation mapping using a custom built model-chooser: comparison of regression tree analysis and multivariate adaptive regression splines. In: Proceedings of the Fourth International Conference on Integrating GIS and Environmental Modeling: Problems, Prospects and Research Needs, Banff, Alberta, Canada, September 2–8, 2000.
Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9, 181–199.
Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn. 1, 81–106.
Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufman Publishers, San Mateo, CA.
R Development Core Team, 2005. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org.
Ridgeway, G., 1999. The state of boosting. Comp. Sci. Stat. 31, 172–181.
Rollins, M.G., Keane, R.E., Zhu, Z. Executive Summary. In: The, LANDFIRE, Prototype Project: nationally consistent, locally relevant geospatial data, tools for wildland fire, management, Rollins, M.G. (Technical Ed.). USDA Forest Service, Rocky Mountain Research Station, Missoula Fire Sciences Laboratory. RMRS-GTR.
Schwarz, M., Zimmermann, N.E., 2005. A new GLM-based method for mapping tree cover continuous fields using regional MODIS reflectance data. Remote Sens. Environ. 95, 428–443.
Sheperd, K.D., Palm, C.A., Gachengo, C.N., Vanlauwe, B., 2003. Rapid characterization of organic resource quality for soil and livestock management in tropical agroecosystems using near-infrared spectroscopy. Agron. J. 95, 1314–1322.
Thuiller, W., 2003. BIOMOD—optimising predictions of species distributions and projecting potential future shifts under global change. Global Change Biol. 9 (10), 1353–1362.
Tomppo, E., 1991. Satellite image-based national forest inventory of Finland. Int. Arch. Photogramm. Remote Sens. 28, 419–424.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Springer, New York.
Vogelmann, J.E., Howard, S.M., Yang, L., Larson, C.R., Wylie, B.K., Van Driel, N., 2001. Completion of the 1990s national land cover data set for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources. Photogram. Eng. Remote S. 67, 650–662.
Wood, S.N., Augustin, N.H., 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecol. Model. 157, 157–177.
Yee, T.W., MacKenzie, M., 2002. Vector generalized additive models in plant ecology. Ecol. Model. 157, 141–156.
Yee, T.W., Mitchell, N.D., 1991. Generalized additive models in plant ecology. J. Veg. Sci. 2, 587–602.